Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England

Nils Gutacker, MSc, Chris Bojke, PhD, Silvio Daidone, PhD, Nancy Devlin, PhD, Andrew Street, PhD

Background. *The English Department of Health has introduced routine collection of patient-reported outcome data for selected surgical procedures to facilitate patient choice and increase hospital accountability. However, using aggregate health outcome scores, such as EQ-5D utilities, for performance assessment purposes causes information loss and raises statistical and normative concerns.* ***Objectives***. *For hip replacement surgery, we explore a) the change in patient-reported outcomes between baseline and follow-up on 5 health dimensions (EQ-5D), b) the extent to which treatment impact varies across hospitals, and c) the extent to which hospital performance on EQ-5D dimensions is correlated with performance on the EQ-5D utility index.* ***Methods***. *We combine information on pre- and postoperative EQ-5D outcomes with routine inpatient data for the financial year 2009–2010. The sample consists of 21,000 patients in 153 hospitals. We employ hierarchical ordered probit risk-adjustment models that recognize the multilevel nature of the data and the response distributions. The treatment impact is modeled as a random coefficient that varies at the hospital level. We obtain hospital-specific empirical Bayes (EB) estimates of this coefficient. We estimate separate models for each EQ-5D dimension and the EQ-5D utility index and analyze correlations of EB estimates across these.* ***Results***. *Hospital treatment is associated with improvements in all EQ-5D dimensions. Variability in treatment impact is most pronounced on the mobility and usual activities dimensions. Conversely, only pain/discomfort and anxiety/depression correlate well with performance measures based on utilities. This leads to different assessments of hospital performance across metrics.* ***Conclusions***. *Our results indicate which hospitals are better than others in improving health across particular EQ-5D dimensions. We demonstrate the importance of evaluating dimensions of the EQ-5D separately for the purposes of hospital performance assessment.* ***Key words:*** *EQ-5D; patient-reported outcomes (PRO); hierarchical ordered probit; provider profiling; quality measurement; performance assessment.* ***(Med Decis Making 2013;33:804–818)***

Recent years have seen a growing trend to measure and publish hospital data on health outcomes to facilitate patient choice and increase provider accountability.[1,2] The focus of these activities has been on measures of mortality, readmission, or adverse events, which are easily derived from clinical records but reveal little about the health of the vast majority of patients. To allow for a more sensitive assessment of hospital performance, it is necessary to move away from a focus on relatively rare ''failure'' outcomes toward more comprehensive and sensitive measures of patients' health outcomes.[3–5]

Since April 2009, all providers of publicly funded inpatient care in the English National Health Service (NHS) have been required to collect both EQ-5D[6] and condition-specific data for 4 elective procedures: unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs.[7] Eligible patients are invited to report their health status before and 3 or 6 months after surgery. The changes in patients' health status are expected to ''provide an indication of the

Received 8 February 2012 from Centre for Health Economics, University of York, UK (NG, CB, SD, AS) and Office of Health Economics, London, UK (ND). The project was funded by the National Institute for Health Research (NIHR) in England under the Health Services Research (HSR) stream (project number 09/2000/47). The views expressed are those of the authors and may not reflect those of the NIHR HSR program or the Department of Health. Revision accepted for publication 17 December 2012.

Address correspondence to Nils Gutacker, Centre for Health Economics, University of York, Alcuin A Block, York, YO10 5DD, UK; e-mail: nils.gutacker@york.ac.uk.

DOI: 10.1177/0272989X13482523

outcomes or quality of care delivered to NHS patients''[7(p5)] and can be analyzed to identify systematic variation across hospital providers with finer granularity than previously possible.

Traditionally, patient-reported outcome (PRO) measures have been collected and analyzed primarily within clinical trials to assess the treatment effect on patients' health. Their application in the context of routine performance assessment on a national scale breaks new ground and requires an appropriate methodology that takes into account the characteristics of the data and their intended use as measures of the relative quality of hospital treatment.[8]

The NHS Information Centre has developed a preliminary risk-adjustment methodology that is currently being applied to the PRO data.[9] For the EQ-5D, this involves transforming the patients' EQ-5D health profiles into utility-weighted index scores and estimating multivariate regression models to relate posttreatment utility scores to the pretreatment scores and case-mix controls. The advantage of this approach is that patient health is expressed in terms of a (quasi-)continuous score, which facilitates statistical analysis and allows for ranking of hospitals with respect to a single performance metric: their ability to influence posttreatment utilities or, equivalently, changes in scores over time. However, for the purposes of performance measurement, identifying best practice, and informing patient choice, the costs of aggregation may outweigh the benefits. We build this argument around 3 points.

First, any form of aggregation causes loss of detail and information.[10] Once constructed, an index measure cannot reveal information about the underlying components and the degree to which hospitals affect these. Certain hospitals may perform well on one EQ-5D dimension but fall short on another. Detailed information on the performance on each dimension can help to identify the source of the problem and foster improvement through adoption of best practice.[10]

Second, the use of an aggregation function introduces exogenous variation that can bias statistical inference and raises normative concerns about whose preferences the weights should reflect.[10,11] In some circumstances, one may be willing to accept the weights underpinning the aggregation function, for example, when conducting economic evaluations of health technologies from a societal perspective.[12] But this is not always justified. The use of aggregate outcome data to inform patient choice raises normative concerns because it imposes a common valuation of health dimensions. In fact, reporting relative hospital performance with respect to risk-adjusted postoperative EQ-5D utility is only justified if all (prospective) patients share the same relative values. But patients may be heterogeneous with respect to their relative valuations of health dimensions or their relative valuations may differ from those of the general public.[13,14] If so, analyzing variation on the level of health dimensions is more appropriate as it allows patients to apply their own values when interpreting performance data.

Third, the use of performance data derived from EQ-5D utility scores may be limited by patients' difficulties in interpreting these quantities. In a recent qualitative study, Hildon and colleagues[15] interviewed patients and clinicians about their views on 4 different metrics of hospital PRO performance, including mean follow-up score, mean change in score, proportion reaching a specified threshold at follow-up, and proportion reaching a minimally important difference. Their results suggest that ''for patients . . . , unlike measures of height or weight, PRO . . . scores are unfamiliar and their values have no immediate meaning. It's therefore necessary to transform them into interpretable forms, or indeed into experiences rather than metrics, to make them useful.''[15(p11)] Furthermore, patients ''could not distinguish between the four [metrics], but liked a percentage, or what was for them intuitive scaling.''[15(p10)] Analyzing responses on EQ-5D dimensions rather than utility scores allows reporting performance in a similar form to the way that the data were originally collected. Hospitals could then be compared with respect to the risk-adjusted probability of a given patient to report, for example, no problems with mobility or pain/discomfort at follow-up.

To explore these claims, we assess hospital performance with respect to self-reported health outcomes for hip replacement patients. We focus on the EQ-5D and develop a multilevel risk-adjustment model for each of the 5 functional dimensions. Our approach draws on the literature on longitudinal modeling[16,17] and on cost-effectiveness in multicenter trials[18] to analyze variation in treatment impact across hospitals. More specifically, we model the hospital-specific contribution to posttreatment EQ-5D response as a random coefficient that varies between hospitals. The empirical Bayes (EB) estimates of this coefficient are then interpreted as capturing relative hospital quality. We assess the correlation between performance assessments on the level of EQ-5D dimensions and aggregated utility scores.

METHODS

Data

Our study exploits EQ-5D data routinely collected from English patients who had a hip replacement during April 2009 to March 2010. All providers of NHS-funded care are required to participate in the survey.[7] This includes all NHS-operated hospitals and private treatment centers. Patients 15 years or older who undergo elective, unilateral hip replacement surgery are invited to take part in the survey.[19] We extract information on each patient's pre- and postoperative EQ-5D health profile and EQ-5D utility score, in which the latter is calculated using the UK time tradeoff (TTO) utility weights.[20] The pretreatment (baseline) survey is collected either during the initial outpatient appointment that precedes hospital admission or at the day of admission. Follow-up data are collected by the NHS Information Centre via postal survey approximately 6 month after surgery. To ensure consistency with respect to the timing of measurements while retaining as much information as possible, we exclude all observations for which the recorded time between baseline survey and admission exceeds 12 weeks or the follow-up period is either shorter than 20 weeks or longer than 1 year.

We link these data to the Hospital Episode Statistics (HES) inpatient database, which contains detailed information on all inpatient care provided in English hospitals. The depth of information contained in HES allows us to account for a wide range of clinical and demographic risk factors. These include the most frequent main diagnoses (e.g., osteoarthritis, rheumatoid arthritis),[21] the weighted Charlson score of comorbidities,[22–24] the number of additionally coded comorbidities, whether it was a primary or revision surgery and whether the revision was due to problems with the existing implant, patient age, sex, and the deprivation profile of the patient's neighborhood of residence.[25–27] We only retain patient records that can be matched to the PRO survey and for which we observe a full EQ-5D profile at baseline and follow-up.

Statistical Modeling

The objective of the empirical analysis is to obtain estimates of the relative systematic impact of hospital providers on patients' posttreatment health outcomes. We estimate hierarchical ordered probit models[28–30] separately for each of the 5 EQ-5D dimensions. We then compare the results with those obtained from a linear regression on the EQ-5D utility scores to study the practical implications of using disaggregated health dimensions for assessment of hospital performance.

Let y^*_{ijt} denote the health status (with respect to, for example, anxiety/depression) of patient $i = 1, \ldots, n_j$ in hospital $j = 1, \ldots, J$ at time point $t \in [0, 1]$. Health status is assumed to be continuous but not directly observable. Instead, we observe patients' own assessment of their status on the 3-point EQ-5D response scale ($m = 1, 2, 3$ with 1 = no problems, 2 = some problems, 3 = extreme problems). The mapping of latent, continuous status y^*_{ijt} to observed, discrete responses y_{ijt} is given by the standard threshold model[31]

$$y_{ijt} = \begin{cases} 3, & if \quad y^*_{ijt} \leq \kappa_1 \\ 2, & if \quad \kappa_1 < y^*_{ijt} \leq \kappa_2 \\ 1, & if \quad y^*_{ijt} > \kappa_2 \end{cases}, \qquad (1)$$

where the threshold parameters κ are unobserved and must be estimated from the data. The categories are ordered from worst to best. This facilitates the qualitative interpretation of regression coefficients, where a positive sign indicates improvements in latent health and, thus, the probability of reporting no problems.

Each patient provides measures of his or her health status pre- and posttreatment. Both responses are outcomes of the same measurement process as well as being (partly) determined by common factors, such as patient characteristics and baseline level of latent health. Our interest lies in the latent health gain that follows from hospital treatment and the degree to which variation in health gain can be systematically associated with the provider of care. We make the assumption that, conditional on baseline health and a set of risk factors, patients do not select into hospitals based on unobservable characteristics and that the health of patients in different hospitals would follow the same trajectory if untreated. This allows us to interpret the variation in latent health gain across hospitals as a measure of relative quality performance.

Our data are characterized by a hierarchical structure, with measurement points clustered in patients, which themselves are clustered in hospitals. Given the nonlinear nature of our model, these data can be analyzed in 2 ways. One can collapse the hierarchy into 2 levels and model posttreatment latent health as a function of lagged, observed (pretreatment) response y_{ij0}, observed patient characteristics, and a hospital effect.[32] Alternatively, one can treat both

pre- and posttreatment latent health as left-hand side variables and estimate longitudinal models with unobserved patient heterogeneity.[16,17,30] We adopt the second approach because it allows us a) to explicitly account for unobserved, time-invariant determinants of latent health; b) to use information contained in both observations to estimate threshold parameters; c) to acknowledge heterogeneity in latent health within a response group as well as random noise in reported pretreatment health; and d) to extend the model in a natural way should more measurement points become available in the future.[33]

Latent health status at any time point t is then described by the outcome equation

$$y^*_{ijt} = \alpha_{ij} + \zeta_j + x'_{ij}\beta + T^*(\nu_j + x'_{ij}\delta) + \varepsilon_{ijt} \quad (2)$$

with

$$\nu_j = \mu + \gamma_j. \quad (3)$$

The vector x_{ij} is a set of patient-level risk adjustment variables that are, in this study, time invariant and assumed to be strictly exogenous. Treatment is modeled as a dummy variable T, which takes a value of 1 if $t = 1$ (posttreatment) and 0 otherwise. The direct effect of treatment on posttreatment health is given by the coefficient ν_j. We also interact T with x_{ij} to allow for differential effects of patient characteristics on health status at baseline and on the effect of treatment.

Unexplained variation is decomposed into 4 variance components: 1) a patient-specific intercept $\alpha_{ij} \sim \mathcal{N}(0, \sigma^2_\alpha)$ that captures unobserved, time-invariant patient heterogeneity in latent health; 2) a hospital-specific, time-invariant intercept $\zeta_j \sim \mathcal{N}(0, \sigma^2_\zeta)$ that addresses hospital clustering; 3) a random coefficient $\gamma_j \sim \mathcal{N}(0, \sigma^2_\gamma)$ that varies between hospitals and describes the systematic hospital effect on posttreatment latent health; and 4) a serially uncorrelated error term $\varepsilon_{ijt} \sim \mathcal{N}(0, 1)$ that leads to the well-known probit specification. Covariance terms between random effects on the same level of the hierarchy are freely estimated, whereas terms across levels are constrained to zero. The variance partition coefficient τ describes the extent to which unexplained variation in posttreatment latent health occurs at the level of the hospital and is calculated as follows[34]:

$$\tau = \frac{\sigma^2_\gamma + 2^*\mathrm{cov}(\gamma, \zeta) + \sigma^2_\zeta}{\sigma^2_\alpha + \sigma^2_\gamma + 2^*\mathrm{cov}(\gamma, \zeta) + \sigma^2_\zeta + \sigma^2_\varepsilon}. \quad (4)$$

Larger values of τ indicate that more variation in posttreatment latent health is attributable to hospital heterogeneity as captured in the hospital-level intercept and the random coefficient on treatment.

For the EQ-5D utility model, we adapt (2) to a linear specification with an identity link function (i.e., $y^*_{ijt} = y_{ijt}$) and $\varepsilon_{ijt} \sim \mathcal{N}(0, \sigma^2_\varepsilon)$.

All ordered probit models are estimated by maximum likelihood using GLLAMM in Stata 11.0 (StataCorp LP, College Station, TX), where the integrals for the random effects are approximated by adaptive quadrature.[35] Threshold parameters and the scale of the coefficient are identified through constraints on the mean and variance of the error term and the mean of the intercept. The linear EQ-5D utility model is estimated by maximum likelihood using xtmixed in Stata 11.0.

Provider Profiling

Our interest lies in estimates of the relative quality of each hospital, γ_j, captured by the hospital-specific deviation from the average effect of treatment, μ. This parameter is not directly estimated but can be recovered in postestimation using the Bayes theorem with variance estimates plugged in for the unknown population parameters, a technique known as empirical Bayes prediction.[36]

For nonlinear models, we describe hospital performance in 2 different ways. First, we rank hospitals according to their impact on latent health status y^*_{ij1}. This can be directly inferred from $\widehat{\gamma}_j$, where more positive values indicate better performance. Second, we compute the probability of reporting a specific posttreatment outcome ($m = 1, 2, 3$), based on the estimated quality effort of the hospital. For the average patient treated in a hospital of average patient intake, this is given by

$$\mathrm{Prob}\left(y_{j1} = m | \overline{x}, \widehat{\gamma}_j, \widehat{\alpha_{ij}} = \widehat{\zeta}_j = 0\right) = \Phi(\kappa_m - S_{j1}) - \Phi(\kappa_{m-1} - S_{j1}), \quad (5)$$

where

$$S_{j1} = \widehat{\mu} + \overline{x}'\widehat{\beta} + \overline{x}'\widehat{\delta} + \widehat{\gamma}_j \quad (6)$$

and $\kappa_0 = -\infty$, $\kappa_3 = +\infty$. We calculate 95% credible intervals around $\widehat{\gamma}_j$ based on their posterior distribution. Note that these credible intervals are only appropriate for single comparison against a given quantity, such as the average, but are too wide for direct comparisons of specific hospitals.[37] Because our interest is on profiling hospital performance with respect to treatment impact, we do not consider uncertainty in other parameters estimates when calculating credible intervals for $\mathrm{Prob}(y_{j1} = m)$.

Table 1 Descriptive Statistics of Patient Characteristics ($N = 21,565$)

Variable	Description	Mean or %	SD	Min	Max
male	= 1, if patient is male	40.7%			
age	Patient's age in years	68.213	10.465	15	95
wcharlson	Weighted Charlson index of comorbidities	0.307	0.647	0	8
add_comorbidities	Number of additional comorbidities not included in Charlson index	1.934	1.859	0	21
deprivation	Index of Multiple Deprivation, income domain	0.124	0.095	0.010	0.830
pretest	Time between preoperative assessment and admission (in days)	19.223	18.720	0	84
posttest	Follow-up (in days)	205.061	26.912	140	365
Primary surgery					
osteoarthritis	= 1, if main diagnosis is osteoarthritis (*ICD-10*: M15-19)	86.5%			
rheumatoid_arthritis	= 1, if main diagnosis is rheumatoid arthritis (*ICD-10*: M05-06)	0.5%			
other_maindiag	= 1, if main diagnosis is not OA or RA	5.8%			
Revision surgery					
revision_complications	= 1, if revision surgery because of complications with existing implant (*ICD-10*: T84)	6.3%			
revisions_other	= 1, if revision surgery for other reasons	1.0%			

ICD-10, International Classification of Diseases, Tenth Revision; OA, osteoarthritis; RA, rheumatoid arthritis.

Both methods produce identical rankings of relative hospital performance. However, only the second method relates the result back to the original scale of the PRO survey instrument and allows differences across hospitals to be investigated in terms of the probability of achieving a specific health outcome.

RESULTS

Descriptive Statistics and Transition Matrices

Our sample consists of 21,565 patients treated in 153 NHS and private hospitals. The number of patients in each hospital ranges from 1 to 1106 (mean [SD], 140 [124]). We present descriptive statistics of patient characteristics in Table 1.

Elective hip replacement surgery is performed predominantly on elderly patients (mean [SD] age, 68.2 [10.5] years), with osteoarthritis being the most common reason for surgical intervention. The majority of patients in our sample are female (59.3%) and admitted for primary replacement of the hip joint (92.7%). The median time elapsed between baseline survey and date of admission is 14 days (interquartile range [IQR], 5–28 days). The median follow-up period is 197 days (IQR, 192–209 days).

Table 2 presents the transition matrices for each of the 5 EQ-5D dimensions. Rows report the patients' own classification of their status at baseline, and columns show self-reported status 6 months after surgery. Accordingly, patients in the lower triangle report improvements in health status, whereas those in the upper triangle report deteriorations.

For each of the 5 dimensions, a considerable number of patients report no problems at baseline. This is especially pronounced on the self-care and anxiety/depression dimensions, in which 44.1% and 57.6% of patients fall into this category, respectively. Of the patients, 6.3% report no problems prior to treatment with respect to mobility, whereas nearly all patients report at least moderate problems with pain/discomfort (99.1%). Sixty-eight patients report having no problems in any of the EQ-5D dimensions.

The number of patients improving since treatment varies greatly by the health dimension under consideration. The dimension most improved since treatment is pain/discomfort, in which 72.3% of the patients report improvements as indicated by a transition to a more favorable category. In contrast, only 29.7% of patients report improvements on the anxiety/depression dimension.

Figure 1 present the empirical distribution of the EQ-5D utility scores pre- and postintervention. The mean preintervention score is 0.349, and the mean postoperative score is 0.761. Both distributions exhibit typical characteristics of empirical EQ-5D distribution observed for a wide range of medical conditions, including multimodality, discontinuity, and clustering at 1 (''full health'').[38,39] Of the

Table 2 Transition Matrices for All EQ-5D Dimensions

	Posttreatment			
Pretreatment	**No (= 1)**	**Some (= 2)**	**Extreme (= 3)**	**Total**
Mobility				
I have *no* problems in walking about (= 1)	1122	240	0	1362
I have *some* problems in walking about (= 2)	10,621	9470	13	20,104
I am *confined to* bed (= 3)	21	74	4	99
Total	11,764	9784	17	21,565
Self-care				
I have *no* problems with self-care (= 1)	8610	884	13	9507
I have *some* problems washing or dressing myself (= 2)	7637	4074	65	11,776
I am *unable to* wash or dress myself (= 3)	78	152	52	282
Total	16,325	5110	130	21,565
Usual activities				
I have *no* problems with performing my usual activities (= 1)	1003	280	24	1307
I have *some* problems with performing my usual activities (= 2)	8492	7102	420	16,014
I am *unable to* perform my usual activities (= 3)	1367	2351	526	4244
Total	10,862	9733	970	21,565
Pain/discomfort				
I have *no* pain or discomfort (= 1)	152	45	1	198
I have *moderate* pain or discomfort (= 2)	7196	4907	237	12,340
I have *extreme* pain or discomfort (= 3)	3822	4581	624	9027
Total	11,170	9533	862	21,565
Anxiety/depression				
I am *not* anxious or depressed (= 1)	11,449	908	55	12,412
I am *moderately* anxious or depressed (= 2)	5477	2405	187	8069
I am *extremely* anxious or depressed (= 3)	477	450	157	1084
Total	17,403	3763	399	21,565

patients, 87.3% report improvements in health as measured by the EQ-5D utility index, whereas 6.4% report deteriorations.

Regression Results

Table 3 presents parameter estimates and associated standard errors for each of the 5 dimension models and the EQ-5D utility index model.

We find several variables to be associated with self-reported health, both at baseline and follow-up. These include male sex (+), higher weighted Charlson index score (–), number of additional comorbidities (–), and the deprivation profile of the patient's neighborhood of residence (–). Patients admitted for primary surgery tend to report worse health status than those returning for revision surgery related to complications with their existing implant, but this effect is only statistically significant for mobility and pain/discomfort. Similarly, patients with a diagnosis of rheumatoid arthritis tend to report lower levels of health, but the effect is insignificant for the mobility and anxiety/depression dimensions.

The mean effect of treatment on posttreatment latent health is positive and significant for all dimensions, resulting in substantial increases in the probability of reporting no problems after surgery (Table 4). The number of comorbidities and the indicators for revision surgery are negatively associated with the treatment effect, indicating that treatment is less beneficial for multimorbid or revision patients. Similarly, patients living in more deprived areas experience, on average, less improvement in latent health than those residing in less deprived areas. Longer follow-up is also associated with a smaller increase in postoperative latent health, albeit the effect being small. For example, for a patient of average characteristics, the probability of reporting no problems on anxiety/depression is estimated to reduce by 0.3% per additional week of follow-up. Postoperative EQ-5D utility scores are expected to reduce by 0.0027 per additional week of follow-up.

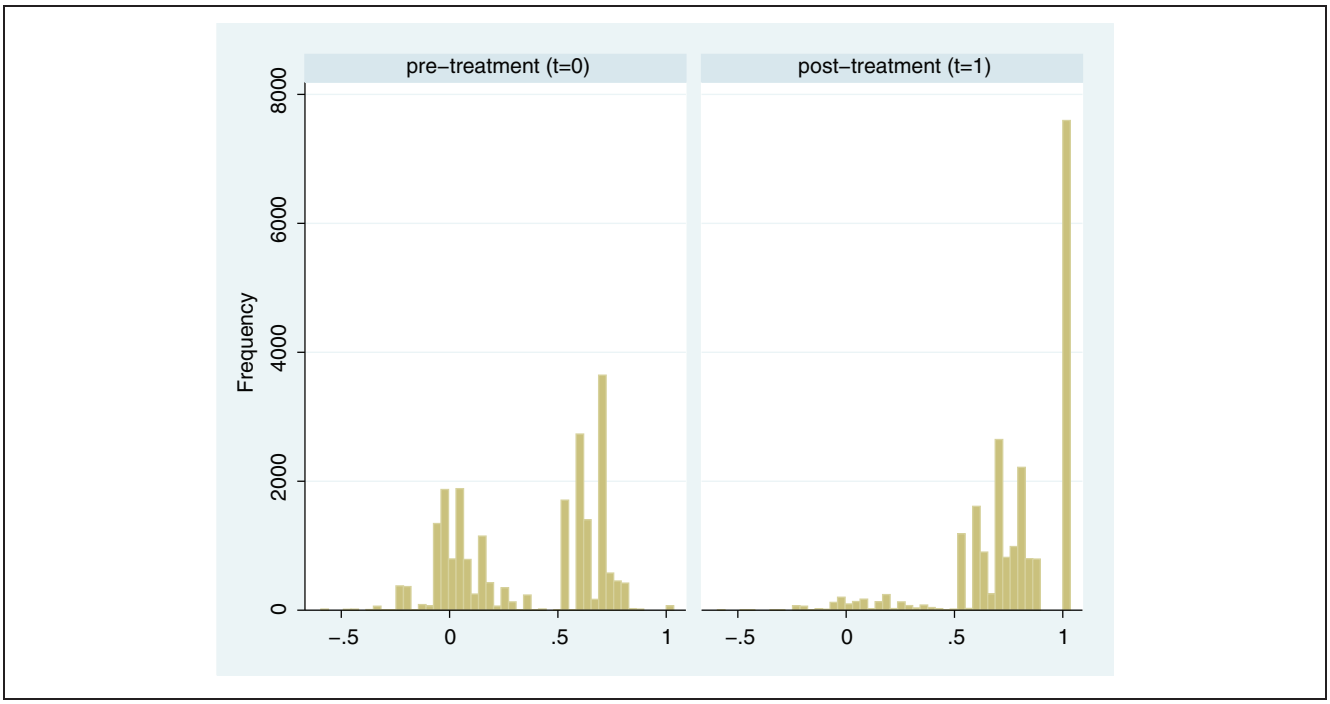


Figure 1 Distribution of EQ-5D utility scores pre- and posttreatment.

All variance components are statistically significant at the 95% confidence level as confirmed by likelihood ratio tests. In contrast, only the covariance term in the EQ-5D utility model is statistically significant. About 1.0% (anxiety/depression) to 4.7% (mobility) of the unexplained variation in latent health is estimated to be associated with the hospital itself.

Assessment of Hospital Performance

Performance on Individual EQ-5D Dimensions and EQ-5D Utility Score

Figure 2a–e presents estimates of hospital performance on the latent health scale (left graph) and the probability scale (right graph), where the latter is calculated for the average patient. Figure 2f presents the results of the EQ-5D utility model, where performance is measured directly on the utility scale. Hospitals located to the left side of each graph perform better than those to the right.

The random coefficient is standardized to zero, which represents the expected outcome for a hospital with average case mix. Hospital performance heterogeneity, as represented by the slope of the curve, is most pronounced on the mobility and usual activities dimensions. For the vast majority of hospitals, credible intervals contain zero, but a small number of hospitals have a statistically significantly different treatment impact. Credible intervals on the mobility dimension are wider than on any other dimension. This reflects the lesser amount of information contained in the data, with only 2 outcome categories being reasonably well populated.

Hospital heterogeneity on the latent health scale translates into differences with respect to hospital-specific probabilities of reporting a given posttreatment health status. The expected probabilities of reporting no problems on the usual activities dimension 6 month after surgery range from 35.8% to 61.8% (calculated for the average patient). In contrast, expected probabilities for the same outcome on the self-care dimension are significantly less dispersed and consistently above 80% for all hospitals. The probability of reporting extreme problems after surgery is close to zero for all models. We refrain from reporting credible intervals around these predicted probabilities to improve the readability of the graphs.

Association of Performance Estimates on EQ-5D Dimensions and the EQ-5D Utility Index

We explore the global agreement between estimates of hospital performance based on individual EQ-

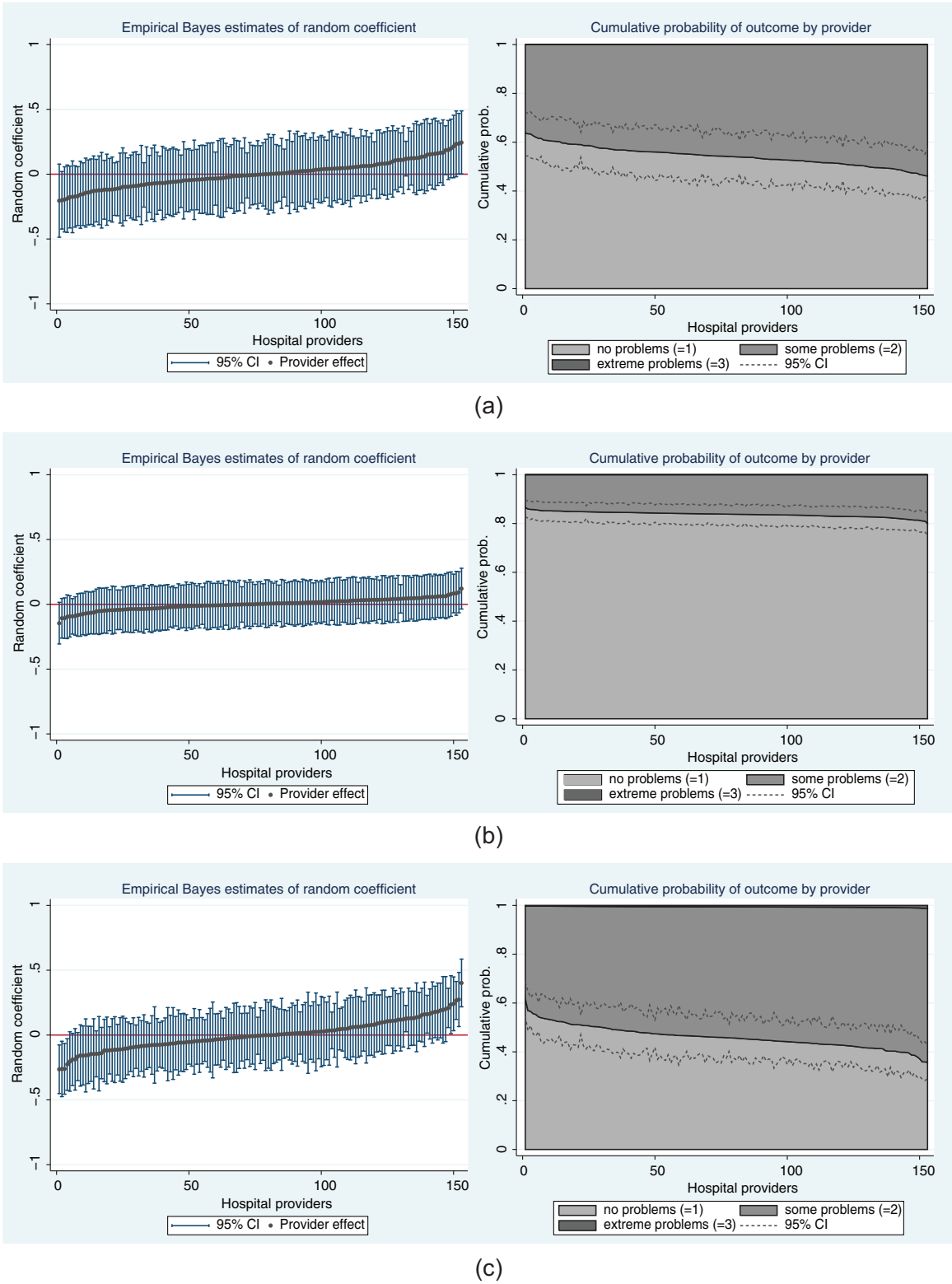
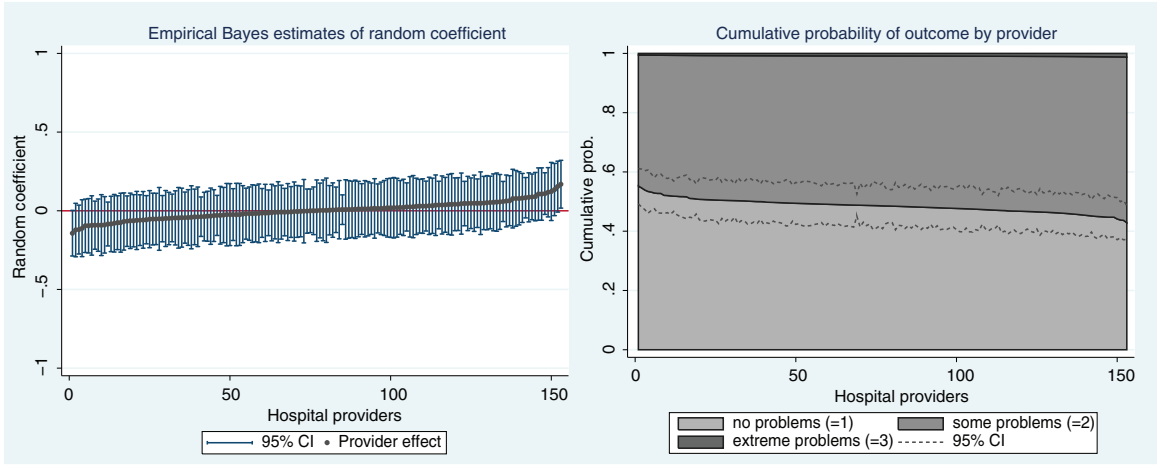
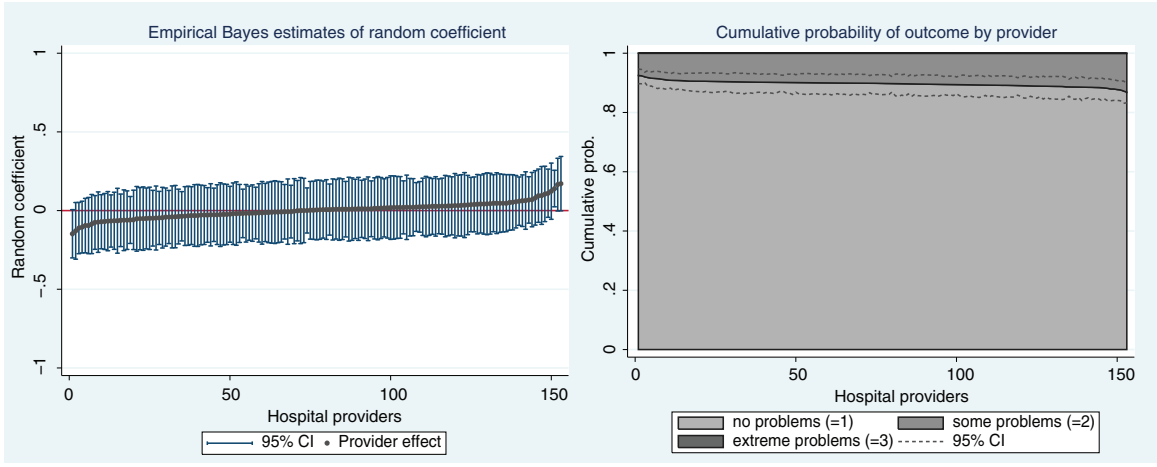


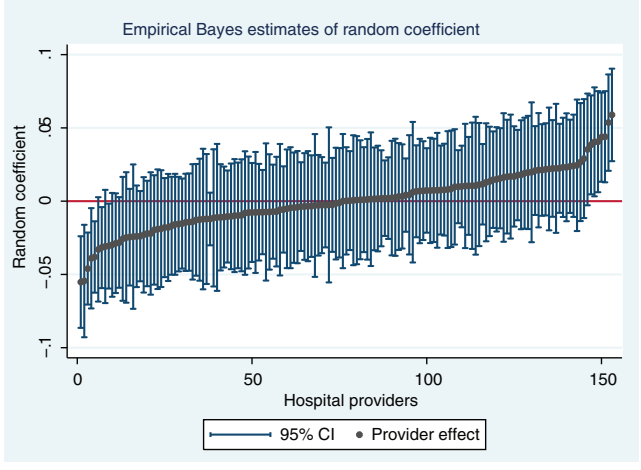
Figure 2 (continued)



(d)



(e)



(f)

Figure 2 Performance estimates on the latent health and outcome scale: (a) Mobility, (b) Self-Care, (c) Usual Activities, (d) Pain/Discomfort, (e) Anxiety/Depression, (f) EQ-5D Utility Index. CI, confidence interval.

Table 3 Regression Results

	EQ-5D Dimensions											
	Mobility		Self-Care		Usual Activities		Pain/Discomfort		Anxiety/Depression		EQ-5D Utility Index	
Variable	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
male	0.172	0.032***	−0.012	0.024	0.091	0.020***	0.302	0.019***	0.458	0.025***	0.072	0.004***
age_15-60	0.022	0.043	−0.139	0.033***	−0.049	0.028	−0.093	0.027***	−0.244	0.034***	−0.032	0.005***
age_71-80	−0.090	0.037*	0.038	0.028	−0.026	0.024	0.028	0.022	0.039	0.029	0.005	0.005
age_81-95	0.283	0.061***	−0.316	0.043***	−0.315	0.036***	−0.083	0.034*	−0.074	0.043	−0.045	0.007***
add_comorbidities	−0.085	0.010***	−0.071	0.007***	−0.051	0.006***	−0.062	0.006***	−0.069	0.007***	−0.019	0.001***
revision_complications	0.162	0.062**	−0.036	0.049	−0.023	0.041	0.247	0.039***	−0.074	0.050	0.013	0.008
revision_other	−0.041	0.167	0.026	0.123	−0.053	0.102	0.148	0.098	−0.264	0.121*	0.004	0.020
deprivation	−0.729	0.174***	−1.128	0.129***	−0.375	0.107***	−1.154	0.104***	−1.056	0.128***	−0.315	0.021***
wcharlson	−0.104	0.027***	−0.146	0.019***	−0.113	0.016***	−0.104	0.015***	−0.100	0.019***	−0.034	0.003***
rheumatoid_arthritis	−0.474	0.243	−0.756	0.167***	−0.382	0.138**	−0.347	0.138*	−0.194	0.164	−0.118	0.027***
other_maindiag	−0.094	0.071	−0.155	0.054**	−0.142	0.045**	−0.049	0.043	−0.081	0.054	−0.030	0.009***
pretest	0.000	0.001	0.003	0.001***	0.001	0.000	0.000	0.000	0.001	0.001*	0.000	0.000*
treatment	2.892	0.101***	1.746	0.103***	2.302	0.084***	2.684	0.081***	1.760	0.106***	0.513	0.015***
treatment × male	0.078	0.036*	0.042	0.030	0.175	0.026***	−0.074	0.025**	−0.163	0.031***	−0.037	0.005***
treatment × age_15-60	−0.038	0.049	−0.033	0.042	−0.089	0.036*	−0.008	0.035	−0.093	0.042*	0.003	0.007
treatment × age_71-80	−0.036	0.043	−0.064	0.036	−0.160	0.030***	0.010	0.029	0.030	0.037	−0.008	0.006
treatment × age_81-95	−0.252	0.068***	0.011	0.052	−0.226	0.045***	0.138	0.044**	0.064	0.054	0.004	0.008
treatment × add_comorbidities	−0.032	0.011**	−0.029	0.008***	−0.054	0.007***	−0.011	0.007	−0.033	0.008***	0.000	0.001
treatment × revision_complications	−0.795	0.072***	−0.537	0.058***	−0.574	0.052***	−0.679	0.050***	−0.353	0.058***	−0.115	0.010***
treatment × revision_other	−0.465	0.190*	−0.830	0.143***	−0.470	0.129***	−0.609	0.126***	−0.444	0.137**	−0.103	0.024***
treatment × deprivation	−0.812	0.197***	−0.524	0.154***	−0.969	0.135***	−0.139	0.132	−0.647	0.152***	0.022	0.025
treatment × wcharlson	−0.113	0.030***	−0.080	0.022***	−0.058	0.020**	−0.020	0.019	−0.020	0.023	0.000	0.004
treatment × rheumatoid_arthritis	−0.531	0.276	−0.325	0.192	−0.344	0.173*	−0.359	0.174*	−0.034	0.198	−0.042	0.033
treatment × other_maindiag	−0.006	0.081	0.039	0.066	−0.015	0.056	−0.071	0.055	−0.032	0.066	0.003	0.011
treatment × posttest	−0.003	0.000***	−0.002	0.000***	−0.002	0.000***	−0.002	0.000***	−0.003	0.000***	−0.000	0.000***
constant	constraint to 0		constraint to 0		constraint to 0		constraint to 0		constraint to 0		0.406	0.006***
κ_1	−3.474	0.073***	−3.388	0.054***	−1.115	0.031***	−0.334	0.028***	−2.506	0.043***	NA	
κ_2	1.670	0.048***	−0.111	0.036**	1.512	0.032***	2.085	0.033***	−0.432	0.035***	NA	
σ^2_ε	constraint to 1		constraint to 1		constraint to 1		constraint to 1		constraint to 1		0.057	0.001***
σ^2_α	0.516	0.037***	1.003	0.038***	0.433	0.019 ***	0.293	0.016***	1.125	0.040***	0.021	0.001***
σ^2_ζ	0.027	0.008***	0.032	0.006***	0.018	0.004***	0.018	0.004***	0.018	0.005***	0.002	0.000***
σ^2_γ	0.027	0.010***	0.009	0.005*	0.026	0.006 ***	0.011	0.004***	0.011	0.005**	0.001	0.000***
$\mathrm{cov}(\zeta, \gamma)$	−0.006	0.007	−0.003	0.004	−0.002	0.004	−0.004	0.003	−0.004	0.004	−0.001	0.000***
τ	0.047		0.023		0.028		0.027		0.010		0.008	
LogL	−19,396		−26,834		−32,251		−32,766		−27,897		−5661	

Significance of variance and covariance components is ascertained by likelihood ratio tests. $N = 21,565$; $J = 153$. NA, not applicable; SE, standard error.
*$P < 0.05$. **$P < 0.01$. ***$P < 0.001$.

Table 4 Predicted Probabilities of Reporting a Given Health Status for a Patient of Average Characteristics

	No (= 1)			Some (= 2)			Extreme (= 3)		
	$t = 1$	$t = 0$	Change	$t = 1$	$t = 0$	Change	$t = 1$	$t = 0$	Change
Mobility	0.543	0.026	0.517	0.457	0.974	−0.517	0.000	0.001	−0.001
Self-care	0.838	0.412	0.426	0.162	0.587	−0.425	0.000	0.001	−0.001
Usual activities	0.460	0.044	0.416	0.534	0.778	−0.244	0.006	0.178	−0.172
Pain/discomfort	0.485	0.012	0.473	0.506	0.550	−0.044	0.009	0.438	−0.429
Anxiety/depression	0.897	0.615	0.282	0.102	0.376	−0.274	0.000	0.009	−0.009

5D dimensions and the utility-weighted EQ-5D index values by calculating Spearman's rank correlation coefficients (Spearman's ρ) and inspecting correlation patterns visually (Figure 3).

The highest rank correlation is observed between performance estimates on the pain/discomfort dimension and EQ-5D utility index (ρ = 0.496), followed by the anxiety/depression dimension (ρ = 0.311). The rank correlation for all other dimensions and the EQ-5D utility index is smaller ($\rho < 0.2$) and, indeed, not statistically significantly different from zero for the mobility and usual activities dimensions.

To explore whether judgment about individual providers would differ depending on which metric is used to assess performance, we identify providers with statistically significantly above/below-average performance on each metric[40–42] and compare the overlap. In 26 of 153 cases, performance classifications differ across metrics (Table 5).

Eleven hospitals (A–K) are identified as above/below-average performers according to the EQ-5D utility model but do not stand out on any of the 5 EQ-5D dimensions. Eight hospitals (L–S) achieve above-average results with respect to at least 1 dimension of the EQ-5D, but this performance is not reflected in their performance estimate on aggregate utilities. Four hospitals (T–W) fall short of the average benchmark on the usual activities dimension but would not be identified as underperformers in terms of their impact on utilities. The disagreement between performance in terms of EQ-5D utilities and individual dimensions is most apparent in the case of hospital X, where the hospital is classified as a low performer in terms of its impact on utilities but is a high performer with respect to restoring its patients' ability to carry out their usual activities.

DISCUSSION

We set out an analytical strategy to explore patient-level and hospital-level variation in categorical responses within and across dimensions of the EQ-5D. This approach does not require assumptions about how to aggregate across health dimensions and offers insight about which dimensions are particularly affected by hospital heterogeneity. We find heterogeneity in performance to be more pronounced across the mobility and usual activities dimensions and less so for the pain/discomfort, anxiety/depression, and self-care dimensions. Furthermore, we find that performance on the utility scale correlates well only with the anxiety/depression and pain/discomfort dimensions. Incidentally, these are the dimensions that receive the highest weighting in the UK TTO EQ-5D tariff.[20] In contrast, the mobility, usual activities, and self-care dimensions have relatively low weights attached to them, and performance heterogeneity remains undetected when analyzing aggregated EQ-5D utility data.

Policy makers are interested in assessing the change in patient-reported outcomes as a result of treatment. There are various ways that this change can be measured and modeled. Our approach has been to model both pre- and posttreatment health status as outcomes of the same reporting process and to conduct multilevel analysis with measurement points clustered in patients, which themselves are nested in hospitals. We argue that this is the appropriate modeling strategy because it acknowledges the features of the data-generating process, allows for patient heterogeneity with respect to observed and unobserved factors, and makes best use of the available information. The presented methodology is readily applicable to other conditions for which EQ-5D data are collected and, in principle, can be extended to other PRO instruments.

In recognition of the expectation that health outcome data are to be used by an audience unfamiliar with the interpretation of complex statistical results (e.g., patients and their relatives, family doctors, managers), we have suggested an intuitively appealing way of summarizing the differential impact that

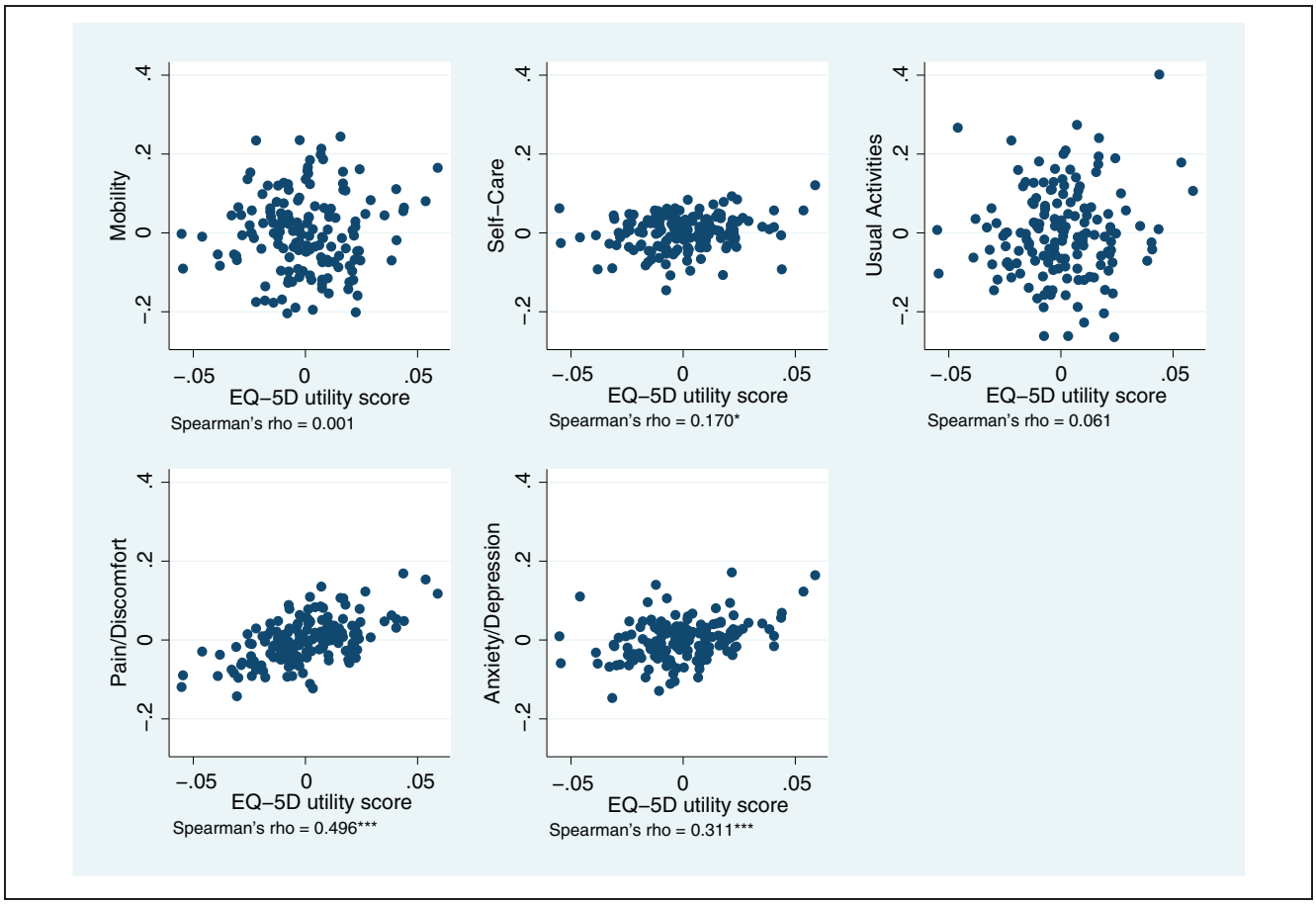


Figure 3 Hospital performance estimates on EQ-5D dimensions and EQ-5D utility scores.

hospitals have on treatment outcomes. Our graphical representation indicates the probability of reporting a given health outcome and shows how these probabilities vary across health dimensions and hospitals. Prospective patients (or their agents) who place greater weight on a particular dimension may use this information to select a hospital that has a differentially greater impact on this than its peers do.

The primary limitation of our proposed approach is the increase in dimensionality of the decision problem for patients. Whereas aggregated scores result in one estimate of hospital performance, our approach generates 5 potentially divergent answers. In a recent study, Dijs-Elsinga and colleagues[43] have shown that a large group of patients favor simple data presentation and prefer one overall measure of hospital quality. But many patients intend to use more detailed quality information when making decisions about where to seek care in the future.[43] The question then arises about how much information should be provided for the different objectives for which performance information can be used (i.e., patient choice, accountability, identification of best practice) and who decides about the relative weighting of each component and objective.[11,44] Our study does not intend to resolve this debate. Rather, we present a means of making inferences about hospital quality and presenting results when health outcomes are assessed through the EQ-5D PRO instrument. How best to communicate such performance data requires careful consideration, to ensure they can be effectively understood and used.

Several issues remain that we have not addressed in this study. First, based on the full information contained in HES, we can identify those patients who have not participated or were not included in the follow-up. We find that, in our data set, only about 50% of eligible hip replacement patients participate in the baseline survey, with a further 8% dropping out of the subsequent survey. These numbers should

Table 5 Examples of Hospitals for Which Performance Assessments Differ across EQ-5D Dimensions and the EQ-5D Utility Model

Hospital	EQ-5D Utilities	Mobility	Self-Care	Usual Activity	Pain/Discomfort	Anxiety/Depression
A	Below	—	—	—	—	—
B	Below	—	—	—	—	—
C	Below	—	—	—	—	—
D	Below	—	—	—	—	—
E	Below	—	—	—	—	—
F	Below	—	—	—	—	—
G	Above	—	—	—	—	—
H	Above	—	—	—	—	—
I	Above	—	—	—	—	—
J	Above	—	—	—	—	—
K	Above	—	—	—	—	—
L	—	Above	—	—	—	—
M	—	Above	—	—	—	—
N	—	—	—	Above	—	—
O	—	—	—	Above	—	Above
P	—	—	—	Above	—	—
Q	—	—	—	Above	—	—
R	—	—	—	Above	—	—
S	—	—	—	Above	—	—
T	—	—	—	Below	—	—
U	—	—	—	Below	—	—
V	—	—	—	Below	—	—
W	—	—	—	Below	—	—
X	Below	—	—	Above	—	—

Hospitals are either statistically *above* or *below* the average or not different from the average (—).

improve in time when data collection procedures become more established. However, falsely assuming that any substantial amount of missing values are generated at random could lead to biased inferences from a nonrepresentative population,[45] raising questions about the validity of the assessment.

Second, in this study, we have controlled for patient risk factors that are deemed clinically relevant, are assumed to be exogenous to the hospital, and can be derived from routine inpatient records. However, we do not claim that this set of control variables is exhaustive: Health outcomes may be affected by nonrandomly distributed, unobserved patient characteristics such as severity of the medical condition or health-related behavior. That said, a strength of our study is that we control for the initial health status with which the patient presents at admission. In many studies, this is unobserved and makes our analysis more robust than possible in the absence of such information.

Third, we do not control for characteristics of the hospital in our analysis, our rationale being that these are within the hospital's control. But they may not be. Hospitals may be constrained in their ability to choose and combine medical resources to their best effect by local regulation, access to factor markets, or, in the short run, the existing capital structure such as age and functionality and whether the hospital operates the service over multiple sites.[46] In this case, the assumption of exchangeability underlying the hierarchical modeling approach may not hold. Furthermore, procedures such as hip replacement are generally followed by extensive physical therapy, which may be delivered outside the hospital. If constraints bind or if quality is not attributable solely to the hospital, our estimates of hospital performance will be biased.

Fourth, our study makes use of a large administrative data set that contains rich information on patient characteristics and the type of care provided. The presented econometric approach is tailored to the data at hand. However, in other countries or disease areas, sample sizes may be smaller or information may be sparse. If patient characteristics are unobserved or cannot be included due to low degrees of freedom, then more of the time-invariant variation between

patients would be captured by the patient random effect. Again, the assumption of exchangeability (i.e., that the unobserved patient heterogeneity is drawn from a random distribution) may become unrealistic and results may be biased.[47] The same argument applies to the random coefficient and the interactions of covariates with the treatment effect. Researchers will need to consider this limitation case by case, based on their data and the available set of risk-adjustment variables.

Finally, further consideration should be given to the role that patient-reported health outcome performance information can play in existing quality assessment frameworks. Although measures of risk-adjusted mortality, readmission, and adverse events have been criticized for their limited granularity and sensitivity,[48] one should not a priori dismiss their ability to identify high- and low-quality providers of care. Further research is required to establish the additional value of outcome data for hospital quality assessments and contrast it to the costs of collection.

ACKNOWLEDGMENTS

We thank Stephen Barasi, Stephen Bloomer, David Nuttall, David Parkin, Aurore Pelissier, Wolfgang Greiner, 3 anonymous referees, and participants of the Health Econometric Data Group seminar series (York), the EuroQoL plenary meeting 2012 (Rotterdam), and the joint CES-HESG Winter conference 2012 (Marseille) for their valuable inputs and comments.

REFERENCES

1. Marshall MN, Shekelle PG, Davies HTO, Smith PC. Public reporting on quality in the United States and the United Kingdom. Health Affairs. 2003;22:134–48.

2. Cutler DM, Ilckman RS, Landrum MB. The role of information in medical markets: an analysis of publicly reported outcomes in cardiac surgery. Am Econ Rev. 2004;94:342–6.

3. Kind P, Williams A. Measuring success in health care—the time has come to do it properly! Health Policy Matters. 2004;9:1–8. Available from: http://www.york.ac.uk/media/healthsciences/documents/research/HPM9final.pdf. Accessed December 1, 2012.

4. Appleby J, Ham C, Imison C, Jennings M. Improving NHS Productivity. London: The King's Fund; 2010.

5. McGrail K, Bryan S, Davis J. Let's all go to the PROM: the case for routine patient-reported outcome measurement in Canadian healthcare. HealthcarePapers. 2012;11:8–18.

6. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37:53–72.

7. Department of Health. Guidance on the Routine Collection of Patient Reported Outcome Measures (PROMs). London: The Stationary Office; 2008. Available from: http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_092625.pdf. Accessed December 1, 2012.

8. Smith PC, Street AD. On the uses of routine patient-reported health outcome data. Health Econ. 2013;22:119–31.

9. Coles J. PROMs Risk Adjustment Methodology—Guide for General Surgery and Orthopaedic Procedures. London: Northgate Informations Solutions Ltd & CHKS Ltd; 2010.

10. Smith PC. Developing composite indicators for assessing health system efficiency. In: OECD, ed. Measuring Up—Improving Health System Performance in OECD Countries. Paris, France: OECD Publications Service; 2002. p 295–316.

11. Parkin D, Rice N, Devlin N. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? Med Decis Making. 2010;30:556–65.

12. Siegel JE, Torrance G, Russell L, Luce B, Weinstein M, Gold M. Guidelines for pharmacoeconomic studies: recommendations from the panel on cost-effectiveness in health and medicine. PharmacoEconomics. 1997;11:159–68.

13. De Wit GA, Busschbach JJV, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? Health Econ. 2000;9:109–26.

14. Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. Health Econ. 2009;18:363–72.

15. Hildon Z, Neuburger J, Allwood D, Van Der Meulen J, Black N. Clinicians' and patients' views of metrics of change derived from patient reported outcome measures (PROMs) for comparing providers' performance of surgery. BMC Health Serv Res. 2012;12:171-71.

16. Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. Springer Series in Statistics. New York: Springer; 2005.

17. Hedeker D, Gibbons RD. Longitudinal Data Analysis. Hoboken, NJ: John Wiley; 2006.

18. Manca A, Lambert PC, Sculpher M, Rice N. Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. Med Decis Making. 2007;27:471–90.

19. NHS Information Centre. A Guide to PROMs Methodology. London: NHS Information Centre; 2010.

20. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.

21. Singh JA. Epidemiology of knee and hip arthroplasty: a systematic review. Open Orthop J. 2011;5:80–85.

22. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40:373–83.

23. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol. 2004;57:1288–94.

24. Bjorgul K, Novicoff W, Saleh K. Evaluating comorbidities in total hip and knee arthroplasty: available instruments. J Orthop Traumatol. 2010;11:203–9.

25. Office of the Deputy Prime Minister. The English Indices of Deprivation 2004 (revised). London: Office of the Deputy Prime Minister; 2004.

26. Clement ND, Muzammil A, MacDonald D, Howie CR, Biant LC. Socioeconomic status affects the early outcome of total hip replacement. J Bone Joint Surg Br. 2011;93:464–9.

27. Neuburger J, Hutchings A, Black N, van der Meulen JH. Socioeconomic differences in patient-reported outcomes after a hip or knee replacement in the English National Health Service. J Public Health (Oxf). 2013;35:115–24.

28. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88:9–25.

29. Gibbons RD, Hedeker D. Random effects probit and logistic regression models for three-level data. Biometrics. 1997;53: 1527–37.

30. Greene WH, Hensher DA. Modeling Ordered Choices. Cambridge, UK: Cambridge University Press; 2010.

31. McKelvey RD, Zavoina W. A statistical model for the analysis of ordinal level dependent variables. J Math Sociol. 1975;4: 103–20.

32. Contoyannis P, Jones AM, Rice N. The dynamics of health in the British Household Panel Survey. J Appl Econometrics. 2004; 19:473–503.

33. National Joint Registry. NJR PROMs questionnaires. 2011. Available from: http://www.njrcentre.org.uk/njrcentre/tabid/199/Default.aspx. Accessed December 1, 2012.

34. Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. Understanding Stat. 2002;1:223–31.

35. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. Stata J. 2002;2:1–21.

36. Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models. J R Stat Soc A. 2009;172:659–87.

37. Goldstein H, Healy MJR. The graphical presentation of a collection of means. J R Stat Soc A. 1995;158:175–7.

38. Basu A, Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. Med Decis Making. 2012;32:56–69.

39. Hernández Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. Value Health. 2012;15:550–61.

40. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. Stat Med. 1994; 13:889–903.

41. Laudicella M, Olsen KR, Street A. Examining cost variation across hospital departments: a two-stage multi-level approach using patient-level data. Soc Sci Med. 2010;71:1872–81.

42. Racz MJ, Sedransk J. Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. J Am Stat Assoc. 2010;105:48–58.

43. Dijs-Elsinga J, Otten W, Versluijs MM, et al. Choosing a hospital for surgery: the importance of information on quality of care. Med Decis Making. 2010;30:544–55.

44. Steyerberg E, Lingsma H. Complexities in quality of care information. Med Decis Making. 2010;30:529–30.

45. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley; 1987.

46. Street A, Scheller-Kreinsen D, Geissler A, Busse R. Determinants of Hospital Costs and Performance Variation: Methods, Models and Variables for the EuroDRG Project. Berlin, Germany: Technical University of Berlin; 2010.

47. Hausman J. Specification tests in econometrics. Econometrica. 1978;46:1251–71.

48. Lilford R, Pronovost P. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. BMJ. 2010;340:955–7.